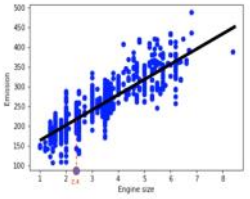
2. Regression

24 Ekim 2022 Pazartesi 10:55

- Very Fast
- No parameter tuning
- Easy to understand,
  And highly interpretable

Simple Linear Regression

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

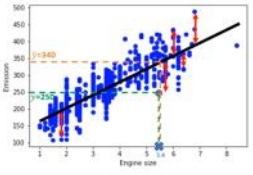


x_1 = 5.4 independent variable
y= 250 actual Co2 emission of x1

$\hat{y} = \theta_0 + \theta_1 x_1$
$\hat{y}$ = 340 the predicted emission of x1

Error = y- $\hat{y}$
= 250 – 340
= -90

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Estimating the parameters

Her value da tekrar edilecek değerler xi ya da yi şeklinde i ile gösterilir.

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 y

$\hat{y} = \theta_0 + \theta_1 x_1$

$\theta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$\bar{x} = (2.0 + 2.4 + 1.5 + \dots)/9 = 3.03$

$\bar{y} = (196 + 221 + 136 + \dots)/9 = 226.22$

$\theta_1 = 39$

$\theta_0 = \bar{y} - \theta_1 \bar{x}$

$\theta_0 = 226.22 - 39 * 3.03$

$\theta_0 = 125.74$

Intercept

Coefficent

$$\hat{y} = 125.74 + 39x_1$$

Model Evaluation in Regression Models

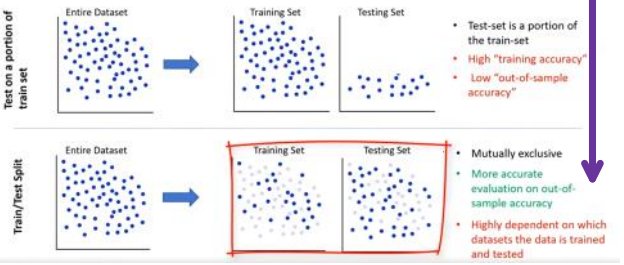
What is training & out-of-sample accuracy?

- **Training Accuracy**
  - High training accuracy isn't necessarily a good thing
  - Result of over-fitting
    - Over-fit: the model is overly trained to the dataset, which may capture noise and produce a non-generalized model
- **Out-of-Sample Accuracy**
  - It's important that our models have a high, out-of-sample accuracy
  - How can we improve out-of-sample accuracy?

Training Accuracy değerinin yüksek olmasını iyi bir şey olarak yorumlamak doğru değildir. Bu değerin yüksek olması over-fit olduğunu gösterir.

Modelin doğruluğu açısından out-of-sample değerinin yüksek olmasını isteriz.

Train/Test split evaluation approach



How to use K-fold cross-validation?

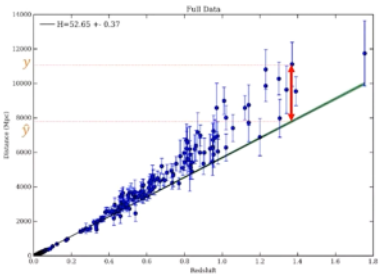
Interpretable - yorumlanabilir

Appropriate - uygun

Decrease - azalmak

remains constant - sabit kalır

What is an error of the model?



$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

Relative absolute error

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

Relative squared error

R-squared $R^2 = 1 - RSE$

MAE (Mean Absolute Error):
- Hataların mutlak değeridir.
- Anlaşılması en kolay olandır.

MSE (Mean Square Error):
- Karesi alınmış hataların ortalamasıdır.
- MAE'den daha popülerdir. Bunu sebebi odak noktasının daha büyük hatalar olmasıdır.
- Hata oranlarını katlayarak gösterir ve daha anlaşılır sonuçlar verir.

RMSE (Root Mean Square Error):
- This is one of the most popular of the evaluation metrics because Root Mean Squared Error is interpretable in the same units as the response vector or Y units, making it easy to relate its information

Multiple Linear Regression

Predicting continuous values with multiple linear regression

$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \dots$

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$

$\hat{y} = \theta^T X$

$\theta^T = [\theta_0, \theta_1, \theta_2, \dots]$ $X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \end{bmatrix}$

X: Independent variable Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Using MSE to expose the errors in the model

$\hat{y} = \theta^T X$

$\hat{y}_i = 140$ the predicted emission of x_i

$y_i = 196$ actual value of x_i

$y_i - \hat{y}_i = 196 - 140 = 56$ residual error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

Estimating multiple linear regression parameters

Making predictions with multiple linear regression

Estimating multiple linear regression parameters

- How to estimate θ?
  - Ordinary Least Squares
    - Linear algebra operations
    - Takes a long time for large datasets (10K+ rows)
  - An optimization algorithm
    - Gradient Descent
    - Proper approach if you have a very large dataset

Making predictions with multiple linear regression

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$\hat{y} = \theta^T X$

$\theta^T = [125, 6.2, 14, \dots]$

$\hat{y} = 125 + 6.2x_1 + 14x_2 +$

$Co2Em = 125 + 6.2EngSize + 14\ \text{Cylinders} + \dots$

$Co2Em = 125 + 6.2 \times 2.4 + 14 \times 4 + \dots$

$Co2Em = 214.1$

Regression algorithms

- Ordinal regression
- Poisson regression
- Fast forest quantile regression
- Linear, Polynomial, Lasso, Stepwise, Ridge regression
- Bayesian linear regression
- Neural network regression
- Decision forest regression
- Boosted decision tree regression
- KNN (K-nearest neighbors)